

De la proportion analogique en apprentissage

Arnaud Delhay-Lorrain *, **Laurent Miclet ****,

*** IUT de Lannion - Université de Rennes 1 - IRISA/CORDIAL*

Département Informatique, Rue Edouard Branly - BP 30219, 22302 Lannion Cedex

** ENSSAT - Université de Rennes 1 - IRISA/CORDIAL*

Rue de Kerampont - BP 80518, 22305 Lannion Cedex

{Arnaud.Delhay,Laurent.Miclet}@univ-rennes1.fr

Sections de rattachement : 27

Secteur : Secondaire

***Résumé.** Cet article s'intéresse à la proportion analogique, une forme simple du raisonnement par analogie, et décrit son utilisation en apprentissage artificiel. Nous nous attachons plus particulièrement à définir une nouvelle notion, la dissimilarité analogique et à l'appliquer à des séquences. Après avoir défini la proportion analogique, la dissimilarité analogique et la résolution approchée d'équations analogiques, nous décrivons deux algorithmes qui rendent opérationnels ces notions pour des objets numériques ou symboliques et pour des séquences de ces objets. Ces notions et ces algorithmes ont montré leur intérêt pour une application de classification d'objets, et pour une autre dans la génération de nouveaux exemples (par résolution approchée d'équations analogiques) pour aider un système de reconnaissance de caractères manuscrits à s'adapter très rapidement à un nouveau scripteur. Une discussion générale sur l'apport du raisonnement par analogie pour l'apprentissage artificiel termine cet article.*

***Mots-clés.** Apprentissage artificiel, classification, analogie.*

1. Introduction

Cet article se propose d'appliquer certaines notions du raisonnement par analogie dans le domaine de l'apprentissage supervisé non paramétrique. Ce sujet est exposé en détail dans MICLET et al. (2008). En apprentissage supervisé, on dispose d'un ensemble d'apprentissage \mathcal{S} composé d'objets, chacun étant associé à sa supervision (une étiquette de classe, par exemple). La question est de savoir à quelle classe on doit affecter un nouvel objet, à partir de la seule connaissance de l'ensemble d'apprentissage (CORNUÉJOLS and MICLET (2002)).

Le principe de l'apprentissage non paramétrique est de ne faire aucune hypothèse sur la distribution statistique des classes. La technique la plus simple de ce domaine est celle de l'apprentissage par *plus proche voisin* : quand arrive un objet extérieur à \mathcal{S} , on lui attribue la classe de l'élément le plus ressemblant dans \mathcal{S} .

Dans la même famille, la technique de l'apprentissage par analogie fait appel à un argument plus sophistiqué. Donnons-en un exemple intuitif sur des objets qui sont représentés par des séquences de lettres. Soit la séquence **cherchera**, dont on veut déterminer la classe; supposons que dans l'ensemble d'apprentissage se trouvent les trois séquences : **recommencer**, **commencera**, **rechercher**, avec respectivement pour classes *infinitif*, *futur*, *infinitif*. On attribuera à **cherchera** la classe *futur*, par un raisonnement qui s'énonce comme ceci. Puisque, dans l'univers des séquences

recommencer est à commencera comme rechercher est à cherchera

la supervision de **cherchera** est donc la solution de l'équation sur les classes :

infinitif est à futur comme infinitif est à x

d'où :

$x = \textit{futur}$.

Ce petit exemple fait apparaître deux notions de base : d'abord celle de *proportion analogique entre quatre objets* (des séquences de lettres dans l'exemple ci-dessus) qui s'exprime sous la forme "*A est à B comme C est à D*". Il faut bien sûr donner une signification à « comme » et à « est à ». En effet, dans un autre exemple, la proportion

“jument est à poulain comme vache est à veau”

porte seulement sur la sémantique des mots, pas sur leur morphologie.

Nous verrons que cette notion peut s'étendre à une proportion approchée, que l'on pourrait exprimer par "*A est à B à peu près comme C est à D*". Par exemple :

“jument est à poulain à peu près comme vache est à bufflon”
“recommencer est à commencera à peu près comme rechercher est à chercherai”

Nous donnerons une quantification et une méthode de calcul du terme « à peu près ».

L'autre notion que cet exemple introduit est celle d'*équation analogique* : si l'on connaît trois termes d'une proportion analogique, peut-on calculer le quatrième? Et peut-on calculer tous les termes qui sont « à peu près » en proportion analogique avec les trois premiers? Nous donnerons des algorithmes pour résoudre ce problème.

D'une manière générale, la proportion analogique est un cas particulier du *raisonnement par analogie* qui a été longuement décrit et étudié depuis les philosophes grecs; ses applications récentes intéressent en particulier les sciences cognitives (Holyoak (2005)), la linguistique et l'intelligence artificielle. LEPAGE (2003) donne une histoire encyclopédique de ce concept et de ses applications à la science du raisonnement et à la linguistique. La restriction à la proportion analogique, en particulier pour les séquences, a été étudiée d'un point de vue méthodologique et algorithmique en particulier dans MITCHELL (1993), HOFSTADTER and the Fluid Analogies Research Group (1994), DASTANI et al. (2003), SCHMID et al. (2003),

YVON et al. (2004). Un autre domaine de l'Intelligence Artificielle est naturellement relié au raisonnement par analogie : celui du raisonnement à partir de cas (CBR ou RàPC). Dans le livre de référence de AAMODT and PLAZA (1994), ces deux notions sont confondues. Nous reviendrons dans la conclusion sur la distinction entre CBR et apprentissage par proportion analogique et sur la légitimité de ce dernier.

2. Proportion analogique et équations analogiques

2.1. Les axiomes de la proportion analogique

Il n'y a pas de définition générale d'une proportion analogique "A est à B comme C est à D" entre quatre objets pris dans un ensemble X , les relations "est à" et "comme" dépendant de la nature de X . Toutefois, d'après la signification usuelle du mot "analogie" en philosophie et en linguistique, trois axiomes de bases sont généralement requis (LEPAGE and ANDO (1996)) :

Definition 1 (Proportion analogique.) Une proportion analogique sur X est une relation sur X^4 . Quand $(A, B, C, D) \in \mathcal{A}$, les quatre éléments A, B, C et D sont dits en proportion analogique, ce qui s'écrit $A : B :: C : D$ et se lit "A est à B comme C est à D". Pour chaque 4-uplet en proportion analogique, les trois axiomes suivants sont requis :

$$\begin{aligned} \text{Symétrie de la relation "comme" : } & A : B :: C : D \Leftrightarrow C : D :: A : B \\ \text{Echange des médians : } & A : B :: C : D \Leftrightarrow A : C :: B : D \\ \text{Déterminisme : } & A : A :: B : x \Rightarrow x = B \end{aligned}$$

2.2. Equations analogiques

Résoudre une équation analogique consiste à trouver le quatrième terme d'une proportion analogique, les trois premiers étant connus.

Definition 2 (Equation analogique.) D est une solution à l'équation analogique $A : B :: C : x$ ssi $A : B :: C : D$.

Selon la nature des objets et la définition des relations, une équation analogique peut ne pas avoir de solution, avoir une solution unique ou avoir plusieurs solutions.

2.3. Proportions analogiques sur \mathbb{R}^n et sur $\{0, 1\}^n$

Quand on se place dans \mathbb{R}^n , il est simple de remarquer que quatre objets sont en proportion analogique quand ils forment un parallélogramme, ce qui peut s'écrire :

$$a : b :: c : d \Leftrightarrow \vec{Oa} + \vec{Od} = \vec{Ob} + \vec{Oc} \Leftrightarrow \vec{ab} = \vec{cd}$$

Il faut aussi de remarquer que quatre objets binaires vérifient les axiomes de la proportion analogique soit quand ils ont tous les quatre la même valeur 0 ou 1, soit quand deux valent 1 et deux valent 0 (sauf dans les cas (0,1,1,0) et (1,0,0,1)). Il y a donc au total 6 quadruplets binaires en analogie parmi les 16 possibles. Quand les objets sont des éléments de $\{0,1\}^n$, il suffit que chaque quadruplet de coordonnées vérifie l'une des 6 proportions analogiques précédentes pour que les objets vérifient les axiomes de la proportion analogique. Ces définitions sont détaillées dans MICLET et al. (2008).

2.4. Proportions analogiques sur les séquences

Nous utilisons maintenant les notions usuelles de la théorie des langages : alphabet Σ , mot (ou séquence) sur Σ^* , longueur d'un mot, mot vide ϵ , facteur, sous-séquence. Par exemple, sur l'alphabet $\Sigma = \{a, b\}$, la séquence $aabbaa$ est un élément de Σ^* de longueur $|aabbaa| = 6$, $bbaa$ en est un facteur et aba en est une sous-séquence.

Nous ajoutons un nouveau symbole à Σ , que nous notons \smile , pour obtenir un alphabet augmenté Σ' . Son interprétation est celle d'un symbole "vide" nécessaire pour les sections qui suivent.

Definition 3 (Equivalence sémantique.) *Soit x une séquence de Σ^* et y une séquence sur Σ'^* . Les séquences x et y sont sémantiquement équivalentes si la sous-séquence de y composée des lettres de Σ est x . Nous notons cette relation par \equiv . Par exemple : $ab \smile a \smile a \equiv abaa$.*

Un alignement est une correspondance lettre à lettre entre quatre séquences, dans lesquelles des lettres \smile peuvent être insérées de façon à ce qu'elles prennent la même longueur. La correspondance $(\smile, \smile, \smile, \smile)$ n'est pas permise.

Definition 4 (Alignement entre quatre séquences.) *Un alignement entre quatre séquences $u, v, w, x \in \Sigma^*$, est un mot z sur l'alphabet $(\Sigma \cup \{\smile\})^4 \setminus \{(\smile, \smile, \smile, \smile)\}$ dont la projection sur la première, la seconde, la troisième et la quatrième composante sont sémantiquement équivalentes à u, v, w and x .*

Nous supposons qu'il existe une relation de proportion analogique dans Σ' , c'est-à-dire que pour chaque quadruplet a, b, c, d dans Σ' , la relation $a : b :: c : d$ vaut soit *VRAI* soit *FAUX*. Nous proposons maintenant de définir la proportion analogique entre quatre séquences d'objets en utilisant à la fois la proportion analogique entre les objets qui les composent et l'alignement entre les quatre séquences.

Definition 5 (Proportion analogique entre séquences.) *Soit u, v, w and x quatre séquences de Σ , sur lequel existe une proportion analogique. Nous disons que u, v, w et x sont en proportion analogique s'il existe quatre séquences u', v', w' and x' de même longueur dans Σ' , avec les propriétés suivantes :*

1. $u' \equiv u, v' \equiv v, w' \equiv w$ et $x' \equiv x$.
2. $\forall i \in [1, n]$ sont des proportions analogiques $u'_i : v'_i :: w'_i : x'_i$ dans Σ' .

Par exemple, soit $\Sigma' = \{a, b, \alpha, \beta, B, C, \sim\}$ avec les proportions analogiques $a : b :: A : B$, $a : \alpha :: b : \beta$ and $A : \alpha :: B : \beta$. L'alignement suivant entre les quatres séquences $aBA, abBA, ba$ et βba est une proportion analogique sur Σ^* :

a	\sim	B	A
α	b	B	A
b	\sim	a	\sim
β	b	a	\sim

3. Dissimilarité analogique

Nous proposons dans cette section de généraliser la relation de proportion analogique entre quatre objets ou quatre séquences d'objets en introduisant la notion de *dissimilarité analogique* (DA). Pour des raisons méthodologiques et opérationnelles, il est souhaitable que la DA vérifie les propriétés suivantes :

Cohérence analogique. $DA(u, v, w, x) = 0 \Leftrightarrow u : v :: w : x$

Symétrie de "comme". $DA(u, v, w, x) = DA(w, x, u, v)$

Echange des médians. $DA(u, v, w, x) = DA(u, w, v, x)$

Inégalité triangulaire. $DA(u, v, z, t) \leq DA(u, v, w, x) + DA(w, x, z, t)$

Asymétrie de "est à". En général, $DA(u, v, w, x) \neq DA(v, u, w, x)$

3.1. Dissimilarité analogique entre objets de \mathbb{R}^n et de $\{0, 1\}^n$

Une DA dans \mathbb{R}^n se définit intuitivement par $DA(a, b, c, d) = \delta(d, e)$, où e est le quatrième sommet du parallélogramme construit sur a, b et c et où δ est une distance dans \mathbb{R}^n . Il est alors aisé de vérifier que les propriétés précédentes sont vraies.

Pour quatre objets binaires, nous savons depuis le paragraphe 2.3 qu'il y a six cas de proportions analogiques exactes sur les 16 quadruplets binaires possibles. En affectant la valeur $DA(a, b, c, d) = 0$ à ces six quadruplets, la valeur 1 à tous ceux qui ont trois 0 et un 1 (ou le contraire) et la valeur 2 aux quadruplets $(0, 1, 1, 0)$ et $(1, 0, 0, 1)$, on démontre que les propriétés ci-dessus sont vérifiées.

Pour des objets de $\{0, 1\}^n$, additionner les valeurs précédentes sur les coordonnées (puis diviser par n si l'on veut normaliser) produit également une dissimilarité analogique ayant les quatre propriétés désirées. C'est cette définition que nous utiliserons dans la suite.

3.2. Dissimilarité analogique entre séquences : définition et calcul.

Definition 6 (Dissimilarité analogique entre séquences.) *Le coût d'un alignement entre quatre séquences est la somme des dissimilarités analogiques entre les quadruplets de lettres définis par cet alignement. La dissimilarité analogique entre quatre séquences est le coût de l'alignement le moins coûteux entre les quatre séquences.*

La définition précédente permet de calculer la dissimilarité analogique $DA(u, v, w, x)$ avec un algorithme de programmation dynamique (appelé SEQUANA4), qui progresse de façon synchrone dans les quatre séquences pour construire un alignement optimal (MICLET et al. (2008)). L'entrée de cet algorithme est un alphabet augmenté Σ' sur lequel une dissimilarité analogique $DA(a, b, c, d)$ a été définie. La sortie est la dissimilarité analogique entre quatre séquences Σ^* , c'est-à-dire $DA(u, v, w, x)$. Cet algorithme a une complexité en temps en $\mathcal{O}(|u|.|v|.|w|.|x|)$.

La validité de SEQUANA4 peut être montrée par récurrence, car il est construit sur le principe de la programmation dynamique. La DA calculée entre les séquences possède les propriétés suivantes : *cohérence avec l'analogie, symétrie du "comme" et l'échange des médians*. La propriété d'*inégalité triangulaire* n'est pas en général assurée.

3.3. Solutions approchées aux équations analogiques

Nous traitons ici du calcul de la solution d'une équation analogique entre séquences, et ce d'une manière étendue : nous donnons en effet le principe d'un algorithme capable de produire toutes les meilleures solutions à ce problème, c'est à dire de construire toutes les séquences en dissimilarité analogique minimale avec trois séquences données.

Definition 7 (Meilleure solution approchée à une équation analogique.) *Soit X un ensemble sur lequel est défini une analogie et une dissimilarité analogique DA . Soit $a : b :: c : x$ une équation analogique dans X . L'ensemble des meilleures solutions approchées à cette équation est donnée par :*

$$\{y : \underset{y \in X}{\text{ArgMin}} DA(a, b, c, y)\}$$

Ce sont donc les objets $y \in X$ qui sont les plus proches d'être en proportion analogique avec a, b et c . Cette définition s'applique au cas de la solution exacte à une équation analogique quand on force la dissimilarité analogique à être nulle.

Nous pouvons facilement élargir ce concept et définir l'ensemble de k -meilleures solutions à une équation analogique $a : b :: c : x$. Informellement, c'est le sous-ensemble de k éléments de X qui ont une DA minimale quand ils sont associés en quatrième position de l'équation avec a, b et c .

Dans \mathbb{R}^n , il n’y a qu’une seule meilleure solution approchée à une équation analogique, qui peut être explicitement calculée. Dans $\{0, 1\}^n$, l’approche naïve est d’examiner chaque élément de l’ensemble et de garder les y qui minimisent $DA(a, b, c, y)$. Dans ce cas la DA possède la propriété d’inégalité triangulaire et nous avons donné un algorithme plus rapide dans MICLET et al. (2008).

Nous décrivons maintenant un algorithme (appelé SOLVANA) qui trouve l’ensemble des meilleures solutions approchées à l’équation $a : b :: c : x$ quand les objets sont des séquences sur un alphabet étendu, sur lequel une DA a été définie.

Cet algorithme utilise la programmation dynamique pour construire un tableau à 3 dimensions. Quand cette construction est terminée, un retour en arrière est fait pour produire un graphe de toutes les meilleures solutions.

L’alignement de quatre séquences de longueurs différentes est réalisé en insérant optimalement des lettres \surd dans les trois séquences connues et en calculant en ligne la quatrième séquence, de telle façon que les quatre séquences aient finalement la même longueur. En parallèle, on cumule les dissimilarités analogiques dans l’alphabet augmenté sur chaque étape de l’alignement des séquences. La complexité algorithmique de cet algorithme est en $O(m * p^3)$, où $m = Card(\Sigma')$ et p est la longueur moyenne des séquences.

Si on cherche non pas l’ensemble des solutions optimales, mais les k meilleures solutions pour k quelconque, cet algorithme n’est plus adapté. Il faut alors employer une technique du type *branch and bound*, comme l’algorithme A^* (en version de base avec une heuristique identiquement nulle ou de manière plus élaborée si l’on peut définir une fonction heuristique admissible).

4. Conclusion

En conclusion, cet article a mis en évidence une définition fondée de la proportion analogique et de la dissimilarité analogique, ainsi que leur mise en forme algorithmique. Ces deux algorithmes peuvent apporter de bons résultats dans deux domaines de l’apprentissage artificiel comme illustré dans MICLET et al. (2008).

En Intelligence Artificielle, le raisonnement par analogie est souvent, comme la citation de l’introduction l’a déjà évoqué, pratiquement confondu avec le CBR. Il s’agit de transférer de l’information d’un domaine bien connu vers un autre, en utilisant le fait que les deux domaines ont une structure commune (GENTNER et al. (2001)). Dans le cas particulier de la proportion analogique, les deux domaines sont confondus. Cette restriction a l’inconvénient de limiter les ambitions (en particulier sur la véracité cognitive du modèle), mais en revanche permet comme on l’a vu de donner un cadre formel et algorithmique fondé.

Pour en rester aux problèmes d’apprentissage de règles de classification, la question centrale est de savoir s’il est plausible que des objets en proportion analogique aient obligatoirement des classes également en proportion analogique. C’est en effet sur cette hypothèse que notre étude est fondée. Si l’on s’en tient, comme nous l’avons fait, à des proportions analogiques triviales dans les classes, le raisonne-

ment est expérimentalement valable, comme le montre l'expérience de classification d'objets binaires de l'article MICLET et al. (2008). D'autre part, une thèse sur des données structurées en arbre a été commencée. Elle pourrait s'appliquer, par exemple, à l'apprentissage de la prosodie en synthèse de la parole.

Remerciements : Le premier auteur remercie tous les membres de l'équipe CORDIAL et ceux du Département Informatique de l'IUT de Lannion pour leur soutien, sans lequel son intégration recherche n'aurait pu se réaliser.

Bibliographie

- MICLET, L., S. BAYOUDH, and A. DELHAY, 2008. Analogical dissimilarity : Definition, algorithms and two experiments in machine learning. *JAIR* 32.
- CORNUÉJOLS, A., and L. MICLET, 2002. Apprentissage artificiel : concepts et algorithmes. Eyrolles, Paris.
- Holyoak, K. J., 2005. Analogy. *In* The Cambridge Handbook of Thinking and Reasoning. Cambridge University Press., 117–142.
- LEPAGE, Y., 2003. De l'analogie rendant compte de la commutation en linguistique. Grenoble. Habilitation à diriger les recherches.
- MITCHELL, M., 1993. Analogy-Making as Perception. MIT Press, Cambridge, MA.
- HOFSTADTER, D., and the Fluid Analogies Research Group, 1994. Fluid Concepts and Creative Analogies. Basic Books, New York.
- DASTANI, M., B. INDURKHYA, and R. SCHA, 2003. Analogical projection in pattern perception. *Journal of Experimental and Theoretical Artificial Intelligence* 15.
- SCHMID, U., H. GUST, K.-U. KÜHNBERGER, and J. Burghardt, 2003. An algebraic framework for solving proportional and predictive analogies. *In* R. Y. F. Schmalhofer, and G. Katz, editors, Proceedings of the European Conference on Cognitive Science (EuroCogSci 2003). Lawrence Erlbaum, Osnabrück, Germany, 295–300.
- YVON, F., N. STROPPA, A. DELHAY, and L. MICLET, 2004. Solving analogical equations on words. Technical Report ENST2004D005, École Nationale Supérieure des Télécommunications.
- AAMODT, A., and E. PLAZA, 1994. Case-based reasoning : Foundational issues, methodological variations, and system approaches. *Artificial Intelligence Communications* 7 :39–59.
- LEPAGE, Y., and S. ANDO, 1996. Saussurian analogy : a theoretical account and its application. *In* Proceedings of COLING-96. København, 717–722. <http://www.slt.atr.co.jp/~lepage/ps/coling96.ps.gz>.
- GENTNER, D., K. J. HOLYOAK, and B. KOKINOV, 2001. The analogical mind : Perspectives from cognitive science. MIT Press, Cambridge, MA.